Perceived shocks and impulse responses*

Raffaella Giacomini‡, Jason Lu§, Katja Smetanina¶

July 26, 2025

Abstract

This paper develops a novel approach that leverages the information contained in expectations datasets to derive empirical measures of beliefs regarding economic shocks and their dynamic effects. Utilizing a panel of expectation *revisions* for a single variable across multiple horizons, we implement a time-varying factor model to nonparametrically estimate the latent shocks and their associated impulse responses at every point in time. The method is designed to accommodate small sample sizes and relies on weak assumptions, requiring no explicit modeling of expectations or assumptions about agents' forecasting models, information sets, or rationality. Our empirical application to consensus inflation expectations identifies a single perceived shock that closely aligns with observed inflation surprises. The time-varying impulse responses indicate a significant decline in the perceived persistence of the effect of this shock, suggesting that inflation expectations have become more "anchored" over time.

JEL classification: C38, C14, E37, E65

Keywords: Beliefs; Time-varying factor models; Nonparametric estimation; Principal Components Analysis; Heteroskedasticity; Small samples.

*We thank: Gadi Barlevy, Martin Ellison, Guido Lorenzoni, Morten Ravn, Esther Ruiz, Dacheng Xiu, Andrei Zeleneev and seminar participants at several seminars and conferences for useful comments and suggestions.

‡University College London, Department of Economics. Email: r.giacomini@ucl.ac.uk

§International Monetary Fund. Email: jlu2@imf.org

¶University of Chicago Booth School of Business. Email: E.Smetanina@chicagobooth.edu

1 Introduction

Data on expectations are increasingly recognized as crucial for understanding economic dynamics and the formation of beliefs. This paper contributes to the literature by demonstrating how the horizon dimension in expectations datasets, combined with a focus on expectation *revisions*, can be leveraged to derive empirical measures of key economic quantities: the shocks perceived by agents and their corresponding dynamic effects (the impulse responses).

The necessary data are a balanced panel of expectation revisions for a single variable across multiple future horizons and over time. These revisions can be constructed using expectations datasets that encompass both time-series and horizon dimensions, which are widely available in economics. Examples of such datasets include: surveys of expectations (e.g., Blue Chip Analysts, Survey of Professional Forecasters, I/B/E/S Earning Forecasts, Survey of Firms' Inflation Expectations, University of Michigan Survey of Consumers), market-based expectations (e.g., Treasury Inflation-Protected Securities break-even inflation, inflation swap contracts, currency futures, implied volatility from option contracts), and combinations of surveys and/or market-based expectations (e.g., Cleveland Fed inflation expectations).[1]

At the core of this paper lies a simple yet powerful idea: fitting a factor model with potentially time-varying loadings to the panel of expectation revisions across horizons allows us to recover two latent components at any given time. First, we extract a low-dimensional vector of independent common factors - representing the shocks that drive agents' revisions across all horizons. Second, we extract the corresponding loadings - representing agents' beliefs about the dynamic effects of each shock, which can vary over time. A key challenge arises from the small horizon dimension typical of the data, which introduces small-sample bias in existing nonparametric methods for factor extraction (Principal Components Analysis, or PCA). We prefer nonparametric approaches for their flexibility and minimal assumptions and we develop a novel econometric method tailored to overcome these limitations

[1] A revision is defined as the difference between the expectation of a variable at a future horizon (e.g., inflation in May) generated in the current period (e.g., April) and the expectation of the same variable (inflation in May) produced in the preceding period (e.g., March).

(Time-Varying Heteroskedastic Principal Components, or tvHPCA).

To illustrate one of the many possible uses of our method, we apply it to extract historical perceived shocks and impulse responses related to inflation. We construct a term structure of consensus expectation revisions across various horizons by integrating two data sources: the Blue Chip Analysts for short- and medium-term horizons and the Cleveland Fed expectations for long-term horizons. Our key findings are threefold. First, a single perceived shock drives inflation revisions across all horizons and is highly correlated with inflation surprises, particularly those from Stock and Watson (2007)'s model. However, the extracted impulse responses diverge significantly from this model's predictions. Second, the perceived impulse responses exhibit time-varying shapes, showing a secular decline in the perceived persistence of the effect of the shock. Third, we illustrate that our method can inform key policy questions, such as whether shifts in long-term expectations signal beliefs in persistent effects of the shock - indicative of deanchoring - or merely reflect the perception of a large shock. Subject to the caveats of end-of-sample nonparametric estimation, our results suggest that the large changes in long-term expectations observed in 2022 primarily resulted from a large perceived shock rather than deanchoring.

Fitting a factor model to expectation revisions may appear arbitrary; however, we demonstrate that such a factor structure emerges naturally from various theories of expectation formation. For instance, when a rational agent utilizes a Structural Vector Moving Average (SVMA) model to generate expectations (see, e.g., Plagborg-Møller, 2019), the expectation revisions for a given variable across multiple horizons exhibit a factor structure, with factors and loadings corresponding to the structural shocks and structural impulse responses, respectively.[2] The ability to represent a wide range of structural economic models as an SVMA (e.g., Plagborg-Møller, 2019) highlights the robustness and broad applicability of our approach. Importantly, we sidestep the challenges of directly estimating the SVMA model for expectations by instead focusing on the implied factor model for expectation revisions. We further show that both a dynamic factor model for the target variable and certain theories

[2]Note the potential to extract multiple shocks from expectations of only one of the variables in the system.

of information rigidities (e.g., the noisy information model in Coibion and Gorodnichenko, 2015), also imply a factor structure for expectation revisions. The fact that different theories yield a factor structure in revisions underscores the robustness of our approach. Crucially, by allowing the loadings to vary over time, our method accommodates any evolving parameters in agents' expectation formation without requiring explicit modeling of these dynamics.

This paper analyzes the term structure of expectations revision across horizons. Relative to existing research, such as Aruoba (2020) and Crump et al. (2023), which use term structures of expectations to answer different questions, we shift the focus from modeling expectations in levels, which can be susceptible to misspecification, to expectation revisions.[3] This provides a more robust basis for capturing forecasters' perceived shocks and impulse responses across various expectation formation theories. Moreover, we build on Aruoba's idea of combining expectations data from different sources in our application, but use it to address a distinct question.

Our approach also distinguishes itself from studies that use factor models and expectations data to study forecaster disagreement. For example, Herbst and Winkler (2021) estimate dynamic factor models on individual forecasts across multiple variables to understand the factor structure of disagreement, while Kim (2023) quantifies the sources of forecaster disagreement using a noisy information factor model. In contrast, our method recovers shocks and perceived impulse responses from revisions for a single variable across various horizons, thus operating under weaker assumptions regarding expectation formation or rationality. Importantly, because our method is a finite sample approach, it could in principle be applied to individual forecasts as well, provided they have a horizon dimension. This potential application opens up new avenues for gaining insights into disagreement, although we do not explore this direction in the current paper and leave it for future research.

This paper contributes to the empirical macro literature by offering a formal method for measuring beliefs regarding shocks and impulse responses that leverages

[3] In our framework, the term structure in levels can be reconstructed by accumulating revisions from a given starting point, though this process inevitably results in an accumulation of estimation errors.

the horizon dimension of expectations data. Regarding impulse responses, the insights gained from these empirical measures can be valuable across various contexts, and engage with multiple strands of literature. First, beliefs regarding the long-term effects of shocks are critical for central banks, and our method can yield new insights into the anchoring of long-term inflation expectations (e.g. Carvalho et al., 2023) without requiring a model of expectation formation and instead focusing on expectation revisions. Second, our framework can help document new stylized facts related to beliefs about the effects of shocks and the shape of impulse response functions, which can be useful to test theories that link such beliefs to aggregate fluctuations (e.g. Blanchard et al., 2013). Third, our flexible nonparametric approach challenges the conventional dichotomy of permanent versus transitory shocks, allowing for a wider variety of impulse response shapes. This is evidenced by our previously mentioned empirical finding that the impulse response functions for inflation we extract differ markedly from those implied by the permanent-transitory model of Stock and Watson (2007).

Regarding empirical measures of shocks, a key question is whether shocks extracted from expectations data using our method can be given a structural interpretation. Importantly, our approach does not change the core identification strategy; such interpretation depends on additional assumptions - usually about the timing of revisions so that they fall within a narrow window around events like policy announcements (e.g., Kuttner, 2001, Gürkaynak et al., 2005, Gertler and Karadi, 2015, Nakamura and Steinsson, 2018). What distinguishes our contribution is the new perspective it offers to this literature. First, our method provides a rigorous econometric framework for shock extraction that exploits the horizon dimension available in high-frequency expectations data, even if limited. Building on Gürkaynak et al. (2005), which extracts the first principal component of changes in expectations around monetary policy events, our approach can be applied to each expectation measure individually across different horizons without assuming the same shocks drive all measures. This can lead to more accurate estimates of structural shocks by incorporating new data sources, correcting small-sample biases from PCA due to heteroskedasticity, and allowing for time-varying loadings. Second, our method can

recover multiple, independent shocks, satisfying a key criterion for structural interpretation and addressing the literature recognizing multiple shocks in high-frequency data (e.g., Swanson, 2021). Third, our method estimates the impulse responses of each shock, which are inherently distinct. For example, in expectations data for price and quantity, it can differentiate supply shocks (driving prices and quantities in opposite directions) from demand shocks (moving both in the same direction).

The simulations in this paper indicate that the small cross-sectional dimension commonly used in existing studies - such as Swanson (2021)'s analysis, which extracts three principal components from eight variables - may potentially introduce biases when applying standard PCA to extract shocks. Notably, when applying both PCA and our robust tvHPCA method to the framework of Swanson (2021), the extracted shocks were quite similar. Although these results are not reported in the paper, they are available upon request. This suggests that small-sample bias may not have significantly influenced shock extraction in this particular context, but it remains possible that it could be more pronounced in other applications or datasets. Additionally, the method we propose can be useful for extracting shocks from other, new series with a limited horizon dimension.

A key feature of our approach is its foundation on weak assumptions. Focusing on revisions and thus avoiding the need to assume how agents form expectations means that we rely solely on expectations data to study beliefs about shocks and their dynamic effects. This stands in contrast to the extensive literature on testing belief accuracy and rationality, which in addition utilizes data on fundamentals to construct forecast errors and imposes assumptions on belief formation (e.g., the seminal paper by Coibion and Gorodnichenko, 2015 and the literature it spurred). Similarly, studies on "belief distortions" and their impact on aggregate fluctuations (e.g., Bianchi et al., 2022, Enders et al., 2021) also depend on forecast errors, necessitating assumptions about information sets and the connection between expectations and fundamentals.

Another sense in which we rely on weak assumptions is the fact that we adopt a nonparametric approach to estimate both shocks (via PCA) and impulse responses (allowing for smooth and flexible patterns of temporal variation). Our methodology is robust to unmodeled serial correlation in shocks and idiosyncratic errors, and it is tai-

lored to address the small horizon dimension that is typical of expectations datasets. We note however that the core premise of this paper - fitting a time-varying factor model to expectation revisions across horizons - could equally be viewed through a parametric lens, e.g., by considering a classical maximum likelihood state-space model (Anderson and Rubin, 1956) or Bayesian approaches such as Del Negro and Otrok (2008). This would however entail additional modeling choices, including distributional assumptions that may be especially questionable in finite samples, as well as assumptions about serial correlation in factors and idiosyncratic errors, the nature of time variation, and addressing various challenges related to identification and normalization.

Our methodological contribution is the development of a procedure for PCA in factor models with possibly time-varying loadings and a small cross-sectional dimension. Existing approaches (Motta et al., 2011 and Su and Wang, 2017) assume a large cross section, but it is well-documented that PCA performs poorly in small samples due to heteroskedasticity in idiosyncratic errors (e.g., Bai and Wang, 2016). Our tvHPCA method addresses the problem by considering a finite-sample approach to PCA that accommodates heteroskedasticity from the matrix completion literature in statistics (Zhang et al., 2022) and adapting it to allow for time-varying loadings. As with all factor models, the extracted shocks and impulse responses are not separately identified. We address this by normalizing either the shocks or the impulse responses, depending on which of the two objects is the primary interest of the analysis.

The intuition behind tvHPCA is that heteroskedasticity in small samples causes discrepancies between the sample covariance eigenvectors and the true factors, primarily due to the diagonal of the covariance matrix (under the assumption of uncorrelated errors). The solution is to iteratively replace this diagonal with that of a low-rank approximation of the covariance matrix. Implementation applies the algorithm in Zhang et al. (2022) to a nonparametrically estimated local (in time) covariance matrix of revisions. The large time dimension in expectations datasets allows nonparametric estimation of time variation, but the method remains applicable in small samples under the assumption of time-invariant impulse responses by using the global covariance matrix.

The tvHPCA method relies on three main assumptions. First, uncorrelated idiosyncratic errors across horizons are needed for unbiased estimation in finite samples. Although theories of expectation formation suggest a factor structure for the revisions without errors, practical issues such as survey participation variability, rounding and data merging can introduce measurement error, which is plausibly uncorrelated. Importantly, any constant bias in expectations that is common across horizons is differenced out and does not violate this assumption. It is well-known (e.g., Bai, 2003) that PCA is robust to cross-sectional correlation in errors in large samples, but the effects of violation of this assumption may be visible in our small-sample setting. Our simulations show that cross-sectionally correlated errors can bias impulse response estimates and (to a lesser extent) reduce the correlation between estimated and true shocks, however our proposed method continues to outperforms PCA even when the first assumption is violated.

The second key assumption is a stable number of shocks over time and the continuity of time-varying impulse responses, which together facilitate the interpretation of shocks. We determine the number of shocks using a local version of the method by Onatski (2010), which also serves as validation of the stability assumption.[4] Our algorithm then ensures continuity of the impulse responses. We note that our modeling of time variation is common in the nonparametric literature and is based on the assumption of local stationarity. This states that expectation revisions are approximately stationary over short time intervals, thus ruling out immediate changes due to, e.g., new information or policy shocks.[5]

The third key assumption, the standard "incoherence" condition in the matrix completion literature, is similar to assuming strong factors. Simulations reveal that in the worst case - a low signal-to-noise ratio (which can be calculated), many zero

[4]The method by Onatski (2010) is based on large sample theory, similarly to other methods for selecting the number of factors that have been proposed in the literature. Developing a robust procedure for selecting the number of factors for PCA in our finite-sample context remains an open challenge for future research.

[5]Using two-sided kernels for nonparametric estimation potentially raises concerns related to the Lucas critique, as shocks may change agents' behavior and affect local estimates. Our assumption of local stationarity accommodates shock-induced behavioral changes, provided these changes manifest slowly over time, with short bandwidths helping to guard against contamination.

impulse responses, and quickly decaying responses - our ability to recover shocks and the average bias remain stable, although certain impulse response estimates show increased bias.

One of the contributions of Zhang et al. (2022) is the demonstration that HPCA not only performs well in simulations but also has certain theoretical optimality properties. However, these theoretical results depend on the assumptions of time-invariant loadings and serially independent factors and idiosyncratic errors, which appear overly restrictive in our context. Through a series of simulations, we show that such stringent assumptions are in fact not necessary for the tvHPCA method to perform well in realistic small-sample scenarios, aside from some performance deterioration at the sample boundaries (a common issue in nonparametric estimation).

In some applications, confidence intervals for the extracted impulse responses may be valuable, and we outline a bootstrap procedure for this purpose. Since shock extraction itself does not require inference, this step is mainly relevant when the focus is on analyzing impulse responses. However, PCA-based inference is known to perform poorly in small samples (see, e.g., Maldonado and Ruiz, 2021), and there is no established theoretical justification for the bootstrap's validity in this context. Our simulations indicate that bootstrap confidence intervals for impulse responses work well under serially uncorrelated shocks and idiosyncratic errors, though they tend to slightly undercover. The presence of serial correlation in the shocks exacerbates this undercoverage issue. Therefore, caution is warranted when interpreting confidence intervals for perceived impulse responses if extracted shocks are serially correlated. Our nonparametric method favors robustness, but if inference is the main goal, traditional parametric approaches (e.g., Anderson and Rubin, 1956) may provide stronger theoretical guarantees, albeit with strong assumptions that can be particularly difficult to defend in small samples.

The paper is organized as follows. Section 2 outlines the methodology, presenting the factor model idea, its link to theories of expectation formation and the tvHPCA algorithm. Section 3 discusses the simulation results, while Section 4 presents the empirical application. Section 5 concludes. Appendix A provides an in-depth discussion of the assumptions and Appendix B addresses bandwidth selection.

2 Methodology

2.1 Set-up and notation

The information required by our method is a (balanced) panel of expectation revisions for one variable across a term structure of different horizons and over time. To fix notation, denote by t the frequency at which the expectations are measured. At each time $t = 0, \ldots, T$ we assume we have expectations of a target variable (denoted simply by Y_h) for a term structure of horizons $h = 1, \ldots, H$. We allow for flexibility in terms of the frequencies at which the expectations are produced and in the definition of the target variable. The simplest case is when the expectations and the target variable are based on the same frequency (e.g., monthly expectations of monthly inflation) and Y_h is the variable h months ahead, $Y_h = Y_{t+h}$. However, we can also accommodate mixed frequencies (e.g., monthly expectations of quarterly inflation), nowcasting (e.g., the first horizon is current-quarter inflation and t is a month within the quarter) and unequally spaced horizons (e.g., $h = 1, \ldots, H$ could denote 1-, 2- and 8-quarters ahead inflation). Also, Y_h could be measured differently at different horizons (e.g., one horizon could be one-quarter ahead inflation and another horizon could be 5-year 5-year inflation).

Let $\widehat{Y}_{h|t}$ and $\widehat{Y}_{h|t-1}$ denote the expectations of the same target variable Y_h made at times t and $t-1$, respectively. We denote the expectation revision at time t for the target variable at the h-th horizon as $X_{h,t} = \Delta\widehat{Y}_{h|t} = \widehat{Y}_{h|t} - \widehat{Y}_{h|t-1}$.[6] Our panel data is thus given by expectation revisions $X_{h,t}$ for a set of horizons $h = 1, ..., H$ and over time periods $t = 1, ..., T$.

Throughout the paper we use the following notation: for a vector v and matrix M, we denote by v' and M' their transposes and $\|\cdot\|$ denotes the matrix spectral norm.

[6] Some care must be applied to the construction of revisions in the mixed frequency case. E.g., if t is April and the first point in the term structure is current quarter inflation, the expectation at $t-1$ is the expectation for 1-quarter ahead inflation made in March.

2.2 The idea: a factor model of expectation revisions

We model the expectation revisions across horizons h and over time t as a time-varying factor model:

$$X_{h,t} = \lambda'_{h,t} F_t + e_{h,t}, \tag{1}$$

for $t = 1, ..., T$, $h = 1, ..., H$, where $F_t = (F_{1t}, F_{2t}, \ldots, F_{rt})'$ is a vector of $r < H$ independent latent factors, $\lambda_{h,t}$ is a vector of factor loadings for the h-th horizon at time t and $e_{h,t}$ is an idiosyncratic error with variance $\sigma^2_{h,t}$. The model thus assumes that there are a few common, latent factors that drive most of the comovements in expectation revisions across horizons, with loadings (as well as error variances) that are allowed to vary across horizons and over time.

At each time t, we interpret the independent latent factors F_t as the "perceived shocks" and the corresponding loading $\lambda_{h,t}$ for each horizon as the "perceived impulse response" at that horizon, that is, the effect of the corresponding shock on the target variable at the h-th horizon. A perceived impulse response function at time t is the plot of $\lambda_{h,t}$ as a function of h. Intuitively, the perceived shocks represent the drivers of the expectation revisions at time t that were unanticipated at time $t-1$. As discussed in the introduction, giving the shocks a structural interpretation generally requires additional assumptions. However, we note that the factors are independent of each other, so they satisfy one of the requirements for structural shocks. In addition, in some applications it is possible that the information contained in the extracted impulse responses can help give a structural interpretation to the extracted shocks.

Factor models have been extensively applied in economics and finance, see e.g. Chamberlain (1983), Diebold et al. (2005), Stock and Watson (2016), and their econometric properties have been studied by, e.g. Bai (2003), Stock and Watson (2006), Bai et al. (2008), Choi (2012), Bai and Ng (2019). A key challenge in our context is that the number of horizons H, i.e., the cross-sectional dimension in the factor model in (1), is typically small in the datasets of expectations available to economists. A main issue that arises in this context is the plausible presence of heteroskedasticity in the idiosyncratic errors $e_{h,t}$. Under additional assumptions (see e.g. Bai, 2003, Bai, 2009), consistent estimates of both factors and loadings can

be achieved if H is large. A recent literature in statistics has however highlighted that, when H is finite, heteroskedastic errors can lead to inconsistent estimates of the factors and loadings, see e.g. Florescu and Perkins (2016), Zhang et al. (2022). In this paper we adapt the solution proposed in this literature to the general case of time-varying factor models.

2.3 Relationship with theories of expectation formation

This section shows that the factor structure for expectation revisions is compatible with some alternative theories of expectation formation. For simplicity, we assume that the target variable is the variable at time $t+h$, $Y_h = Y_{t+h}$. Additionally, we assume here that the model parameters are constant over time. However, the examples could be adapted to incorporate time-varying parameters, provided the variation is compatible with the assumption of local stationarity for the revisions, as discussed in Section 5. For instance, one could allow for parameters that evolve slowly, ensuring that there is no temporal variation between periods $t-1$ and t. It is easy to see that, for all the theories discussed in this section, this type of time variation in the parameters of the model used to form expectations translates into time-varying loadings in the factor representation for the expectation revisions.

2.3.1 Rational expectations and SVMA model

Suppose a representative agent uses a Structural Vector Moving Average (SVMA) model to forecast the target variable Y_{t+h} (See, e.g., Plagborg-Møller (2019), for how a large class of economic models can be represented in this form). The model for the target variable is thus one of the equations of the SVMA:

$$Y_t = \Theta(L)\varepsilon_t,$$

where ε_t is a vector of structural shocks and $\Theta(L)$ is a lag polynomial (for simplicity here assumed to be of order greater than the number of horizons H) whose coefficients represent the structural impulse responses of the target variable to each structural

shock at the corresponding horizon.

It is easy to see that in this case the expectation revision between times $t-1$ and t for the target variable at the h-th horizon is given by

$$X_{h,t} = \widehat{Y}_{t+h|t} - \widehat{Y}_{t+h|t-1} = \theta_h' \varepsilon_t,$$

where $\widehat{Y}_{t+h|t}$ is the conditional mean implied by the SVMA model. This implies a factor structure for the revisions with no idiosyncratic errors, $X_{h,t} = \lambda_{h,t}' F_t$, where the factors are the structural shocks, $F_t = \varepsilon_t$, and the loadings are the associated structural impulse responses at horizon h, $\lambda_{h,t} = \theta_h$.

We thus see that, if the agent uses a model that can be expressed as a SVMA, our procedure can recover the latent vector of structural shocks and the structural impulse responses. Note that we can in principle recover multiple structural shocks from only observing the expectation revisions for one of the variables in the system (provided the number of horizons is larger than the number of shocks).

2.3.2 Rational expectations and factor model

Factor models are frequently employed to model and forecast macroeconomic and financial variables (see, e.g., Stock and Watson, 2002). For instance, interest rates are typically represented using a factor model that captures the joint dynamics of rates across various maturities (Diebold and Li, 2006). A factor model for the target variable suggests a corresponding factor structure for the expectation revisions. For example, consider the dynamic factor model

$$\begin{aligned} Y_t &= \gamma' \beta_t + v_t \\ \beta_t &= \Phi \beta_{t-1} + \varepsilon_t. \end{aligned} \tag{2}$$

The dynamic Nelson and Siegel model (see, e.g., Diebold and Li, 2006) for a specific interest rate maturity Y_t could be written in this form, with three latent factors in the vector β_t and a specific parameterization for γ that depends on the maturity. Stock and Watson (2007)'s model of inflation is also a special case of (2), with β_t

scalar and $\gamma = \Phi = 1$. The expectation for the target variable at horizon h is

$$Y_{t+h|t} = \gamma' \Phi^h \beta_t,$$

which means that the revision of the expectation for Y_{t+h} made between times $t-1$ and t is given by

$$X_{h,t} = \gamma' \Phi^h (\beta_t - \Phi \beta_{t-1}) = \gamma' \Phi^h \varepsilon_t.$$

We thus once again obtain a factor structure for the revisions with no idiosyncratic errors, $X_{h,t} = \lambda'_{h,t} F_t$, where the factors $F_t = \varepsilon_t$ coincide with the shocks in the state equation specifying the law of motion for β_t and the loadings $\lambda'_{h,t} = \gamma' \Phi^h$ are the impulse responses implied by the model. In this setting, therefore, our method can recover the true impulse responses and can also back out the number of dynamic factors in agents' models by only observing how agents revise expectations for one variable (e.g., one interest rate maturity) across different horizons.

2.3.3 Information rigidities

To see how existing theories of expectation formation with information rigidities could give rise to a factor structure for expectation revisions, consider, e.g., the noisy-information model in Coibion and Gorodnichenko (2015), where the target variable follows an AR(1) process

$$Y_t = \rho Y_{t-1} + \varepsilon_t,$$

with ε_t Gaussian white noise. Here the (single) true shock is ε_t and the true impulse response at horizon h is given by ρ^h. The theory assumes that a representative agent observes a noisy signal of Y_t,

$$Z_t = Y_t + v_t,$$

with v_t a Gaussian white noise process independent of ε_t. The agent forecasts using the Kalman filter, so the expectation for the h-th horizon made at time t is

$$\widehat{Y}_{t+h|t} = \rho^h \widehat{Y}_{t|t},$$

where

$$\widehat{Y}_{t|t} = GZ_t + (1-G)\widehat{Y}_{t|t-1}$$

and G is the Kalman gain, which captures the degree of information rigidity. The revision is then given by

$$X_{h,t} = \widehat{Y}_{t+h|t} - \widehat{Y}_{t+h|t-1} = \rho^h(\widehat{Y}_{t|t} - \widehat{Y}_{t|t-1}).$$

We thus again obtain a factor structure for the expectation revisions with no idiosyncratic errors, $X_{h,t} = \lambda_{h,t}F_t$, where the loadings $\lambda_{h,t} = \rho^h$ correspond to the true impulse responses and the factor is the "filtered shock" $F_t = \widehat{Y}_{t|t} - \widehat{Y}_{t|t-1} = G(Z_t - \widehat{Y}_{t|t-1})$, that is, the surprise from the Kalman filter updating equation that the agent uses to extract the signal from the noise.

2.4 Choosing the number of shocks

Before implementing the algorithm in Section 2.5, one should choose number of shocks, assumed to be constant over time (see Assumption 3(b)). We apply a local version of Onatski (2010) as a guide for this purpose and to verify Assumption 3(b). While other tests from the literature could also be used, all such methods are only justified in large samples. We offer some simulation evidence in Section 3 that Onatski (2010)'s test is capable of recovering multiple shocks, even with a small number of horizons. A rigorous approach to selecting the number of factors in our setting remains however an open question and is left for future research.

The procedure is outlined as follows:

1. Calculate the local eigenvalues from the eigendecomposition of the local covariance matrix, $\psi_{1,t}, \cdots, \psi_{H,t}$,

2. Select r_{max} as a preliminary maximum number of shocks we are interested in testing for (2 in our empirical application),

3. Setting $j = r_{max}+1$, regress $\psi_{j,t}, \cdots, \psi_{j+4,t}$ on a constant and $(j-1)^{2/3}, \cdots, (j+3)^{2/3}$,

4. Set $\delta_t = 2|\hat{\beta}_t|$, where $\hat{\beta}_t$ is the slope coefficient from the above regression,

5. Compute $r(\delta_t) = \max\{i \leq r_{max} : \psi_{i,t} - \psi_{i+1,t} \geq \delta_t\}$, and $r(\delta_t) = 0$ if $\psi_{i,t} - \psi_{i+1,t} < \delta_t$ for all $i \leq r_{max}$

6. If $r_{max} \neq r(\delta_t)$, set $j = r(\delta_t) + 1$ and repeat from step 2 onward, otherwise select $r_t = r(\delta_t)$ as the number of local shocks,

7. Set $r = \max_t\{r_t\}$ to determine the number of shocks in our data across time.

2.5 Estimation of shocks and impulse responses via tvHPCA

This section outlines the estimation of shocks, time-varying impulse responses, and idiosyncratic variances for a specified number of shocks, r. To understand the small-sample challenges we face, consider the tvPCA method discussed by Motta et al. (2011) and Su and Wang (2017). This approach applies PCA to the estimated local covariance matrix rather than the global one. While heteroskedastic errors are not a problem when both the time dimension T and the horizon H are large, inconsistency can arise under heteroskedasticity when H is fixed and small, as highlighted for PCA by Paul (2007), Johnstone and Lu (2009), and Onatski (2012).

To get an intuition for why this happens, note that, under the normalization $\Sigma_{F,t} = I_r$, we have:

$$\Sigma_t = E(X_t X_t') = \Lambda_t \Lambda_t' + \Sigma_{e,t}, \tag{3}$$

where $\Sigma_{e,t}$ exhibits heteroskedasticity. When $H \to \infty$ and the factor is pervasive (i.e. $\Lambda_t \Lambda_t' \to \infty$), then $\Lambda_t \Lambda_t'$ becomes the dominant term in (3). This ensures that the principal eigenvector and eigenvalue of Σ_t are close to those of $\Lambda_t \Lambda_t'$. If H is fixed, however, the first principal component eigenvector will put relatively large weight on

the component of X_t with the largest idiosyncratic variance. Such an eigenvector, however, might be unrelated to any of the columns of Λ_t, which is necessary for identification of factors and loadings.

Assuming that $\Sigma_{e,t}$ is diagonal, heteroskedasticity introduces bias into the diagonal elements of the sample covariance matrix. A common approach to address this issue is to use a diagonal-deletion Singular Value Decomposition (SVD), which involves setting the diagonal elements of the sample covariance matrix to zero before applying the SVD (see e.g. Florescu and Perkins (2016)). However, Zhang et al. (2022) note that this approach can fundamentally alter the singular subspace which determines the factors and their loadings, distancing it from the singular subspace corresponding to the true factors and loadings. Zhang et al. (2022) propose an algorithm, HPCA, for estimating the factor model in the presence of heteroskedasticity when H and T are fixed. The idea is to iteratively impute the diagonal entries of the sample covariance matrix by the diagonals of its low-rank approximation.

The following algorithm extends the procedure in Zhang et al. (2022) to allow for time-varying loadings (i.e., impulse responses).

The tvHPCA algorithm:

The following algorithm delivers estimates of the shocks $\widehat{F}_t$, the time-varying impulse responses $\widehat{\Lambda}_t$, the errors $\widehat{e}_t = X_t - \widehat{\Lambda}_t \widehat{F}_t$, and the diagonal matrix of time-varying idiosyncratic variances $\widehat{\Sigma}_{e,t}$. Define the operator $\Delta(\Sigma) = \Sigma - D(\Sigma)$, where $D(\cdot)$ denotes the diagonal operator that returns the matrix containing only the diagonal entries. For a chosen number of shocks r and for each $t = 1, \ldots, T$, perform the following steps:

A. Preliminary estimation of shocks and impulse responses

Step A1: Obtain a nonparametric estimator of the local covariance matrix of the revisions, henceforth denoted simply by $\widehat{\Sigma}_t$, as:

$$\widehat{\Sigma}_t(u) = T^{-1} \sum_{t=1}^{T} K_b(t/T - u) X_t X_t', \tag{4}$$

where $K_b(\cdot) = K(\cdot/b)/b$ is a kernel function, with bandwidth b. Set maximum number of iterations M.

Step A2: Initialize the local tvHPCA algorithm at $m = 0$ by setting the diagonal entries of $\widehat{\Sigma}_t$ to zero,

$$N_t^{(0)} := \Delta(\hat{\Sigma}_t), \qquad m = 0.$$

Step A3: Perform Singular Value Decomposition (SVD) on $N_t^{(m)}$ and denote its rank-r approximation by $\widetilde{N}_t^{(m)}$. Specifically, for $H \geq r$:

$$N_t^{(m)} = \sum_{i=1}^{H} \psi_{i,t}^{(m)} u_{i,t}^{(m)} (v_{i,t}^m)', \qquad \psi_1^{(m)} \geq \cdots \geq \psi_H^{(m)} \geq 0,$$

$$\widetilde{N}_t^{(m)} = \sum_{i=1}^{r} \psi_{i,t}^{(m)} u_{i,t}^{(m)} (v_{i,t}^m)', \tag{5}$$

where $\psi_{i,t}^{(m)}$ is the i-th largest singular value (i.e., the square root of the eigenvalue) of $N_t^{(m)\prime} N_t^{(m)}$, and $u_{i,t}^{(m)}$ and $v_{i,t}^{(m)}$ are the eigenvectors of $N_t^{(m)} N_t^{(m)\prime}$ and $N_t^{(m)\prime} N_t^{(m)}$, respectively.

Step A4: Update $N_t^{(m+1)} = D(\widetilde{N}_t^{(m)}) + \Delta(N_t^{(m)})$, that is, replace the diagonal entries of $N_t^{(m)}$ by those of $\widetilde{N}_t^{(m)}$:

$$N_{h,j,t}^{(m+1)} = \begin{cases} N_{h,j,t}^{(m)} = \widetilde{N}_{h,j,t}^{(m)}, & h = j; \\ \widehat{\Sigma}_{h,j,t}, & h \neq j. \end{cases}$$

Step A5: Calculate the convergence distance as the maximum change in eigenvalues across horizons and time. Stop if the convergence distance is less than a predefined threshold (we select 10^{-3}),[7] or if $m = M$, otherwise set

[7] A smaller threshold choice will typically lead to more accurate convergence at the cost of slower compute times. Our choice of 10^{-3} can be adjusted depending on the application. In our experience a smaller threshold did not change our estimates but resulted in significantly longer compute times.

$m = m + 1$ and return to Step A3 and continue to iterate.

Step A6: Upon convergence, set $\widetilde{\lambda}_{i,t} = u_{i,t}^{(m)}$ for $i = 1, \ldots, r$ as the estimated normalized impulse responses, which for a given t are identified up to the scale and sign. The estimated common covariance matrix is given by $N_t^{(m+1)}$. The diagonal error covariance matrix estimator is:

$$\widehat{\Sigma}_{e,t} = \widehat{\Sigma}_t - N_t^{(m+1)}.$$

Algorithm A above can be seen as an extension of the original HeteroPCA of Zhang et al. (2022) to account for time-variation in the factor loadings, and it can further be interpreted as the projection gradient descent (PGD) for the following rank-constrained (nonconvex) optimization problem:

$$\min_{rank(\widetilde{N}_t) \leq r} \|\Delta(\widehat{\Sigma}_t - \widetilde{N}_t)\|_F^2, \tag{6}$$

where for a matrix M we write $\|M\|_F = (\sum_{h,j} M_{h,j}^2)^{1/2}$ to denote its Frobenius norm and $\widehat{\Sigma}_t$ and $\widetilde{N}_t$ are defined in (4) and (5) respectively.[8] The algorithm requires choosing a bandwidth b for the estimation of the local sample covariance matrix. Given that the estimation is done via local singular value decomposition followed by several steps of refinements, the bias-variance trade-off necessary to discuss a theoretically optimal bandwidth is non-standard. We therefore follow Su and Wang (2017) and use a data-driven way of selecting an optimal bandwidth by cross-validation, which we describe in the appendix.

Next, to unify the identification of the sign of the impulse responses *across time*, we leverage the fact that the impulse responses are continuous in time. Without loss of generality, we therefore make an additional assumption that impulse responses are on average positive across time[9], which identifies the path of the impulse responses for each shock up to sign. For $t = 2, \ldots, T$, we perform the following steps:

[8] All existing convergence results for PGD do not apply to nonconvex optimization problems such as the one in (6), while Zhang et al. (2022) provide theoretical guarantees for their algorithm.

[9] It is also possible to fix the sign of the shocks's correlation with some external variable instead.

B. Ensuring continuity of impulse responses.

Step B1: For a given shock $k = 1 \ldots, r$, compare the estimated impulse responses at time $t-1$ from step A and assign the sign according to the condition below that ensures continuity across time:

$$\text{sign}(\widetilde{\lambda}_{k,t}) = \begin{cases} \text{sign}(\widetilde{\lambda}_{k,t-1}) & \text{if } ||\widetilde{\lambda}_{k,t} - \widetilde{\lambda}_{k,t-1}|| \leq ||\widetilde{\lambda}_{k,t} + \widetilde{\lambda}_{k,t-1}|| \\ -\text{sign}(\widetilde{\lambda}_{k,t-1}) & \text{otherwise} \end{cases} \tag{7}$$

where $||\cdot||$ is the L^2 norm.

Step B2: Estimate the latent shocks associated with $\widetilde{\Lambda}_t = (\widetilde{\lambda}_{1,t}, \ldots, \widetilde{\lambda}_{H,t})'$ by least squares as

$$\widetilde{F}_t = \widetilde{\Lambda}_t \widetilde{\Lambda}_t' (\widetilde{\Lambda}_t X_t)^{-1}, \tag{8}$$

which will not generally have unit variance per shock across time.

Step B3: For $k = 1, \ldots, r$, estimate the time-varying variance of the k-th shock using a local constant kernel regression [10]:

$$\widetilde{F}_{k,t}^2 = \widehat{\mu}_k + w_{k,t}, \tag{9}$$

to obtain the estimated standard deviation of shock k at time t as

$$\widehat{\sigma}_{k,t} = \max(\sqrt{\widehat{\mu}_k}, 10^{-6}). \tag{10}$$

Step B4(a): The normalization that shocks have unit local variance (Assumption 2(a)) - utilized when the main goal is recovery of the impulse responses - is imposed by scaling the shocks and the associated impulse responses as:

$$\widehat{F}_t = D(\widehat{\sigma}_t)^{-1} \widetilde{F}_t, \qquad \widehat{\Lambda}_t = D(\widehat{\sigma}_t) \widetilde{\Lambda}_t, \tag{11}$$

where $D(\widehat{\sigma}_t)$ is the diagonal matrix of the estimated shock standard devi-

[10] For simplicity, to obtain the nonparametric estimate of the local mean of $F_{k,t}^2$, $\widehat{\mu}_k$, we use the same bandwidth we used for local estimation of the sample covariance matrix in (4).

ations, obtained from Step B3.

Step B4(b): The normalization that shocks have unit effect on all horizons (Assumption 2(b)) - utilized when the main goal is recovery of the shocks - is imposed by setting:

$$\widehat{F}_t = \widetilde{F}_t, \qquad \widehat{\Lambda}_t = \widetilde{\Lambda}_t \tag{12}$$

as the impulse responses are eigenvectors that already have unit L^2- norm.

Step B4(c): The normalization that shocks have unit effect only on the first horizon (Assumption 2(c)) - utilized when the main goal is recovery of shocks that are rescaled to have a unit effect on impact - is imposed by setting:

$$\widehat{F}_t = D(\widetilde{\Lambda}_t[1,:])\widetilde{F}_t, \qquad \widehat{\Lambda}_t = D(\widetilde{\Lambda}_t[1,:])^{-1}\widetilde{\Lambda}_t, \tag{13}$$

where $D(\widetilde{\Lambda}_t[1,:])$ is the first row of Λ_t as a diagonal matrix.

2.6 Confidence intervals for impulse responses

Confidence intervals for impulse responses can be obtained using the bootstrap. The following is a schematic description of a residual-based wild bootstrap procedure that takes into account the uncertainty in both shocks and impulse responses. The algorithm assumes that the shocks and the idiosyncratic errors of the factor model are serially uncorrelated.

1. Apply tvHPCA on the revisions data X_t to obtain the shocks $\widehat{F}_t$, the time-varying impulse responses $\widehat{\Lambda}_t$, the idiosyncratic errors $\widehat{e}_t = X_t - \widehat{\Lambda}_t\widehat{F}_t$ and their diagonal covariance matrix $\widehat{\Sigma}_{e,t}$.

2. Calculate the standardized errors $\widehat{\varepsilon}_t = \widehat{\Sigma}_{e,t}^{-1/2}\widehat{e}_t$.

3. Generate B samples of time points randomly with replacement for each $t = 1, ..., T$ to obtain: $\{t_b\}_{b=1}^{B}$.

4. Use the resampled time points to generate B samples of bootstrapped standardized errors $\widehat{\varepsilon}_t$ and shocks $\widehat{F}_t$ to obtain: $\{\varepsilon_t^b\}_{b=1}^B$, and $\{F_t^b\}_{b=1}^B$.

5. Create B bootstrap samples of data: $\{X_t^b = \widehat{\Lambda}_t F_t^b + \widehat{\Sigma}_{e,t}^{1/2}\varepsilon_t^b\}_{b=1}^B$.

6. Apply tvHPCA on each of the B bootstrap samples and retain the impulse responses estimates: $\{\widehat{\Lambda}_t^b\}_{b=1}^B$.

7. Calculate boostrap confidence intervals for the impulse responses using their empirical distribution across bootstrap samples.

3 Simulations

This section analyses the performance of our method in a similar setting as our empirical application. We first consider time-invariant impulse responses and show that: 1) HPCA is robust to serial correlation in shocks or idiosyncractic errors and outperforms PCA even when the assumption of uncorrelated errors across horizons is violated; 2) weak factors have a lesser impact on the ability to recover shocks and on the average bias of the estimated impulse responses, but result in increased bias for some impulse response estimates; 3) bootstrap confidence intervals for impulse responses tend to undercover in small samples, the more so the higher the serial correlation in the shocks; 4) Onatski (2010)'s method can in principle recover more than one shock even though the number of horizons is small. We then focus on time-varying impulse-responses and show that tvHPCA can recover the time variation, with some deterioration in performance at the boundaries of the sample.

All simulations below are based on 1000 Monte Carlo replications.

3.1 HPCA vs. PCA and robustness to correlation

We generate data from the factor model (1), with $r = 1, H = 7, T = 100$ and equal impulse responses across horizons $\lambda_h = 1$ for all h. We consider the following parameterizations.

- (Independence) Let $F_t \sim i.i.d.N(0,1)$ and $e_{h,t} \sim N(0, \sigma_h^2)$ independent across h and heteroskedasticity: $\sigma_1^2 = 1$, $\sigma_h^2 = 0.25$ for $h = 2, ..., 7$.

- (Serially correlated shocks) Let F_t be an AR(1) process with autoregressive coefficient $\rho = 0.7$ and generate $e_{h,t}$ as in the independent case.

- (Serially correlated errors) Let $F_t \sim i.i.d.N(0,1)$ and generate each $e_{h,t}$ as an AR(1) with autoregressive coefficient $\rho = 0.5$ (rescaling the standard deviation of the AR(1) error by $(1-\rho^2)^{0.5}$ to maintain the same unconditional variance for $e_{h,t}$ as in the independent case).

- (Cross-sectionally correlated errors) Construct the covariance matrix of $e_{h,t}$ as

$$\Sigma_e = \Sigma_{\text{diag}}^{1/2} \left((1-\rho) I + \rho \mathbf{1} \right) \Sigma_{\text{diag}}^{1/2},$$

  where Σ_{diag} is the diagonal matrix containing the same variances of the idiosyncratic errors as in the independent case, I is the identity matrix, $\mathbf{1}$ is a matrix of 1's, and $\rho = 0.1$ (lower correlation) or $\rho = 0.5$ (higher correlation) controls the cross-sectional correlation between errors. This specification preserves the original variances on the diagonal while introducing an equicorrelated off-diagonal structure.

We first investigate the bias in impulse response estimation by reporting in Figure 1 the impulse response estimates λ_1 for the first horizon for all designs (since this is the impulse response most affected by heteroskedasticity in our design). Figure 1 reveals that HPCA improves on the performance of PCA in all cases: it corrects the bias of PCA under independence or serial correlation and it has lower bias than PCA even under cross-sectionally correlated errors (a violation of Assumption 1). The figure also shows that serial correlation does not affect the performance of HPCA, yielding essentially unbiased estimates of impulse responses.

We then assess the ability of HPCA to recover the true factor by reporting in Figure 2 the correlation between the estimated shock and the true shock for the different designs. The figure shows similar conclusions as for the bias: HPCA improves on the

performance of PCA in all designs (it delivers higher average correlation between the estimated and true factor, even under violation of Assumption 1) and is robust to serial correlation (the average correlation between estimated and true factor is not affected by the presence of serial correlation). We observe that cross-sectional correlation in errors has a stronger impact on the bias of the estimated impulse responses - rising by around 20% on average in the highest correlation scenario - than on the accuracy of shock recovery. In this worst-case scenario, the correlation between estimated and true shocks decreases by about 6% on average, but remains nonetheless high at around 0.92.

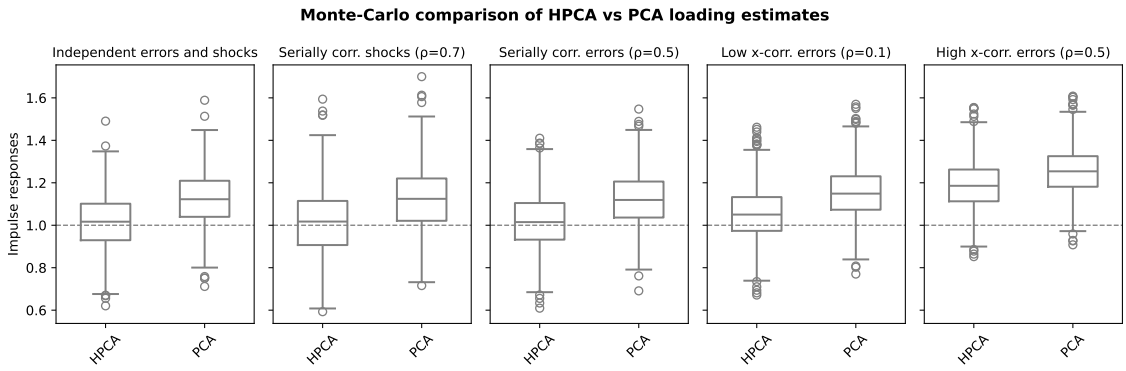

Figure 1: First impulse response estimates: HPCA vs PCA across all DGPs.

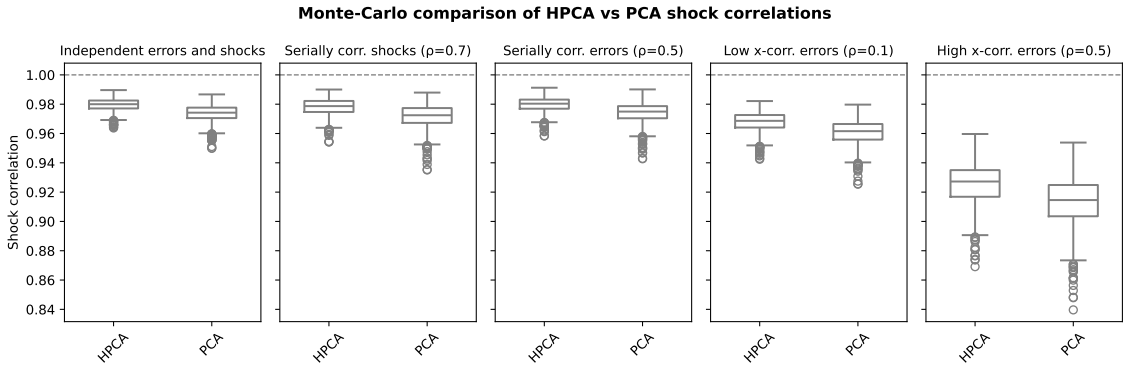

Figure 2: Estimated shock correlation: HPCA vs PCA across all DGPs.

3.2 Impact of weak factors

In this section, we aim to evaluate the implications of violating Assumption 4, by examining the effects of weak factors on our HPCA method. We build on the literature concerning the finite sample performance of PCA by first relating factor strength to the signal-to-noise ratio (SNR), a quantity that can be computed in applications. We then go beyond the existing literature by exploring whether, keeping the SNR constant, the shape of the impulse response functions matters - specifically, their rate of decay to zero and the number of zero responses. We consider the following definition of SNR, which extends the measure considered by Maldonado and Ruiz (2021) to account for heteroskedastic noise:

$$SNR_t = \frac{\sum_{h=1}^{H} \lambda_{h,t}^2}{\max_{h=1}^{H} \sigma_{h,t}^2}, \tag{14}$$

which can be computed in applications by plugging in estimates of the impulse responses $\lambda_{h,t}^2$ and idiosyncratic variances $\sigma_{h,t}^2$. The SNR can be viewed as a way to quantify the amount of information available for estimating the latent shocks and impulse responses. A higher SNR indicates that we can expect more accurate estimates from our HPCA algorithm.

We select Monte Carlo parameters once again to match our application, with $r = 1, H = 7, T = 100$. For simplicity, we set all idiosyncratic variances to 1, so that the SNR is the sum of squared impulse responses. In our empirical application, the SNR varies significantly over time, with the lowest SNR, representing the worst-case scenario, being just above 5. We thus set SNR=5 and vary the shapes of impulse-response functions based on two characteristics: 1) the slope of the non-zero impulse responses (ranging from 0 to 1), and 2) the number of impulse responses that are exactly equal to zero. Figure 3 illustrates two examples of impulse response functions with the same SNR: the left panel depicts a flat response with a slope of 0, while the right panel shows a response that linearly decreases to zero with a slope of 1.

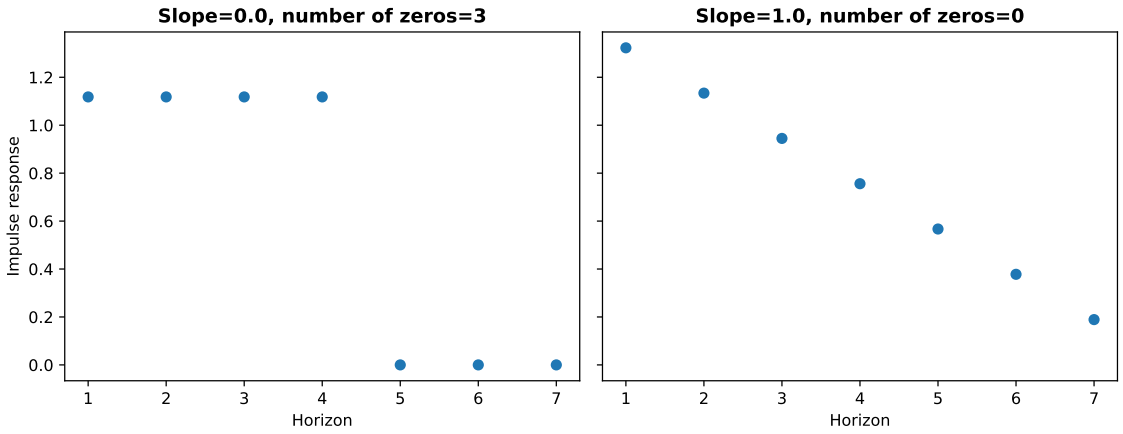

Figure 3: Two examples of impulse response functions with $SNR = 5$.

We assess the performance of HPCA in terms of the the correlation between estimated and true shock, the average (across horizons) impulse response bias, and the maximum (across horizons) impulse response bias. Figure 4 shows the summary of the performance across a matrix of different shapes for the impulse responses, while keeping the SNR fixed at the worst-case scenario of SNR=5.

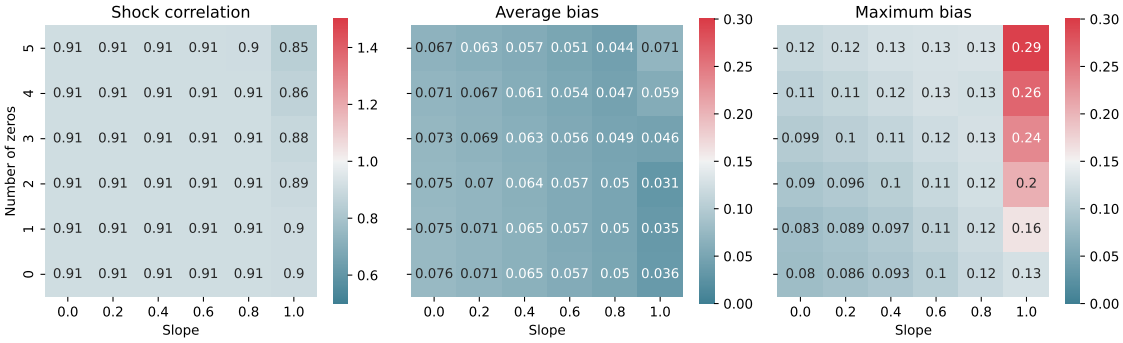

Figure 4: HPCA summary statistics across a matrix of IRF parametrizations.

Overall, HPCA performs well and is largely unaffected by the IRF shape in terms of shock correlation and average bias across horizons. However, maximum bias increases when the IRF has a steep slope and many zeros.

3.3 Bootstrap coverage

We investigate the coverage rates of the bootstrap procedure described in section 2.6 for obtaining confidence intervals for impulse responses. The simulation design is the same as in section 3.1, assuming serially uncorrelated factors and idiosyncratic errors. For each Monte Carlo replication, we generate bootstrap confidence intervals using $B = 1000$ bootstrap iterations. The empirical coverage rates for the confidence interval for the impulse responses across Monte Carlo replications (averaged across horizons) are 93% for a nominal 95%, 87.7% for 90% and 83% for 85%. We thus see a slight tendency to undercover.

We then introduce serial correlation and generate the shock as an AR(1) with autoregressive coefficient 0.7. In this case, the undercoverage of the confidence interval is clearer, with empirical coverages 85.1% for a nominal 95%, 77.9% for 90% and 72.5% for 85%. This is unsurprising, since the bootstrap assumes no serial correlation.

3.4 Ability to recover multiple shocks

We generate data from model (1) with $r = 2, H = 7, T = 100$, with the two shocks in $F_t \sim i.i.d.N(0, 1)$, independent of each other, impulse responses to the two shocks given by $\lambda_h = (1, -(h/H))$, and heteroskedasticity given by $\sigma_h^2 = 0.6 - 0.4(h/H)$, $h = 1, ..., H$. We apply Onatski's (2010) procedure for determining the number of shocks. The procedure correctly identifies the presence of two shocks 82% of the time, and incorrectly identifies the presence of one shock 18% of the time. This indicates that the procedure is in principle able to uncover the presence of more than one shock, even in small samples where the cross-sectional dimension H is very small and there is heteroskedasticity.

3.5 Estimation of time-varying impulse responses

Here we introduce time-varying impulse responses in a model with two shocks and analyze the ability of tvHPCA to recover time-varying impulse responses.

We generate data as in (1), with $H = 7, T = 500, r = 2$. The time-varying impulse responses are $\lambda_{h,t} = [\sin(2\pi(h/H + t/T)), e^{-1}\sin(2\pi((h+2)/H + t/T))]$, such that the second shock explains $1/e$ the variation of the first shock. The time-varying idiosyncratic variances are $\sigma^2_{h,t} = [3 + \sin(2\pi(h/H + t/T))]/5$. The two shocks in F_t are i.i.d. N(0, 1), independent of each other. We note that these parameterizations imply a signal-to-noise ratio SNR (defined in equation (3.2)) for the first factor approximately stable around 5.5, and approximately 0.8 for the second factor. We thus expect the second factor and its impulse responses to be less precisely estimated.

Indeed, we find that both shocks are generally well recovered by the tvHPCA estimation, although the first shock is more precisely estimated (0.94 correlation with the true first shock) than the second shock (0.72 correlation).

Figure 5 shows the nonparametric estimates of the time-varying impulse responses together with pointwise 95% confidence intervals. As expected, the impulse responses to the first shock are more precisely estimated than those to the second shock. The empirical confidence interval coverage rates (averaged across all horizons) are 97.2% and 81.8% for the impulse responses to the first and second shocks, respectively, for a nominal 95% coverage. Figure 5 reveals that our method is generally able to recover the patterns of time variation in impulse responses, with some deterioration in performance at the beginning and end of the sample.

A possible explanation for the deterioration in performance at the sample boundaries is that we use a locally-constant regression in estimating the local covariance matrix, which introduces some bias. This is a known issue with locally-constant estimation, and a common solution is to use locally-linear estimation instead, but this is difficult to apply in our context of estimating covariance matrices as the positive definiteness constraint cannot be easily enforced.[11]

[11] E.g., Chen and Leng (2015) discuss the bias of locally-constant estimation of covariance matrices. They propose a locally-linear estimator, however this method leverages the Cholesky decomposition which is dependent on the order of the variables and is not generally applicable.

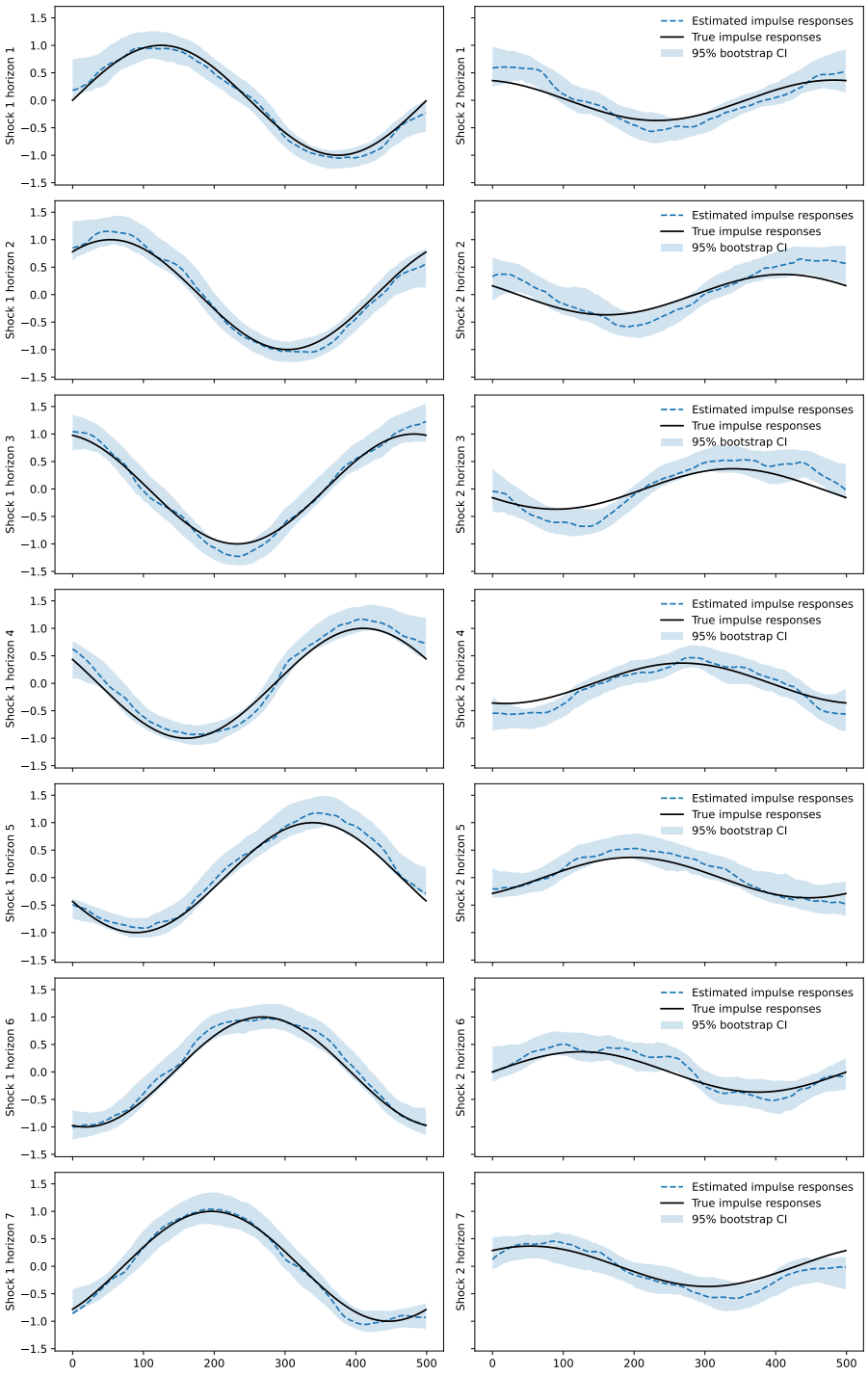

Figure 5: Simulation bootstrap 95% confidence intervals for all impulse responses.

4 Empirical application: perceived shocks and impulse responses of inflation

In this section we use our tvHPCA method to extract historical perceived shocks and impulse response functions from time series of expectations data on inflation.

4.1 Data

Our primary data source is the consensus expectations from the Blue Chip Economic Indicators (2023) (BCEI) survey of professional forecasters. The consensus expectations are the average of the individual expectations across all survey participants at any given point in time. We focus on the expectations of quarterly CPI inflation (annualized rate, percentage). In each calendar month, the survey reports expectations for the quarters of the current and next calendar years. The number of available horizons thus decreases throughout the year and the largest number of horizons available every month is five.[12] From the BCEI we thus consider a balanced panel of monthly expectations for five horizons, for which we use indices $h = 0, ..., 4$ to highlight the fact the first point in the term structure of horizons is a nowcast ($h = 0$) whereas the remaining points are forecasts of the future four quarters ($h = 1, ..., 4$).[13]

In addition to the short- and medium-horizons expectations described above, we would like to include measures of long-term inflation expectations. However, a limitation of the BCEI is that long-term inflation expectations are surveyed less

[12]For instance, a survey conducted in January has eight horizons (2 full years), while a survey conducted in December of the same year has only the current quarter and the four quarters of the next calendar year.

[13]Some attention should be paid to issues of timing and information sets in the BCEI. The survey is usually published in the middle of each calendar month, shortly after the CPI data release. The information set of the forecasters typically contains the previous month's CPI release, although this is not necessarily guaranteed (for instance the forecaster may submit their response prior to the CPI release). One thus needs to establish if the expectation provided in a given month constitutes observed data, a nowcast or a forecast. For example, for the September survey the latest CPI release is that of August and therefore Q3 CPI inflation is not yet observed. This means that the Q3 expectation in September is a nowcast (corresponding to $h = 0$) and the revision is computed as the change relative to the August survey expectation for Q3.

frequently than short- and medium-term expectations.[14]

For long horizons, we therefore opt to use the inflation expectations series published monthly by the Federal Reserve Bank of Cleveland (2023) (CF thereafter). This series is obtained as a model-based composite of BCEI surveys and data (the latest CPI release, treasury yields, and inflation swap prices). We find this to be the most appealing option as one can interpret these expectations as the result of updating the infrequent professional survey expectations for long horizons using high-frequency data. The CF expectations are updated each month immediately after the CPI release; therefore, their timing and information set are aligned with those of the BCEI survey.[15] From the CF inflation expectations series, we consider the expectation revisions of 2-year 3-year and 5-year 5-year CPI inflation.[16] These correspond approximately to medium- and long-term inflation expectations, and they extend the time span captured by the expectations from one year (for the BCEI) to 10 years.[17]

To summarize, by combining expectations data from BCEI and CF we obtain a balanced panel of monthly revisions of CPI inflation expectations from February 1982 to July 2023, for a term structure of seven horizons: the 0th-4th quarterly

[14]Long-term expectations from the Survey of Professional Forecasters (SPF) are also similarly not available every month. There are some market-based measures of inflation expectations available at higher frequencies, including for long horizons. The breakeven inflation rate (the difference between yields on nominal and real U.S. debt of similar maturities), for example, is often quoted as a measure of inflation expectations. However, this measure is not ideal for our analysis because it is confounded with the inflation risk premium.

[15]Note that the data prior to 2009 is constructed ex-post by CF using real-time data.

[16]We infer the CF 5-year 5-year forward inflation expectation (π_{5y5y}) from the 10-year (π_{10y}) and 5-year (π_{5y}) CF inflation expectations (which forecast average inflation over the next 10 and 5 years, respectively) using: $(1+\pi_{5y})^5(1+\pi_{5y5y})^5 = (1+\pi_{10y})^{10}$, and similarly for π_{2y3y}. These transformations ensure that the time periods encompassed by the different horizons for the CF expectations are non-overlapping, as is the case for the BCEI expectations. Overlapping periods would introduce spurious covariance between revisions across horizons and thus contaminate our results.

[17]While we could in principle construct expectations for additional horizons from the CF (e.g., 2-year 1-year, up to 9-year 1-year forward rates), we avoid this for two reasons. First, the CF expectations of inflation typically decay quickly to the long-term trend, so longer horizons would add little information. Second, the inclusion of numerous intermediate horizons would inflate the covariance among revisions at these horizons and possibly bias our estimation toward explaining them.

horizons and the 2-year 3-year and 5-year 5-year horizons.[18]

A possible concern of augmenting the BCEI expectations with the CF long-term expectations is that the results are driven by the CF expectations. Figure 6 provides reassurance that this is not the case, as we see that the local sample covariance matrix from the merged dataset has similar eigenvalues as when considering only BCEI data.

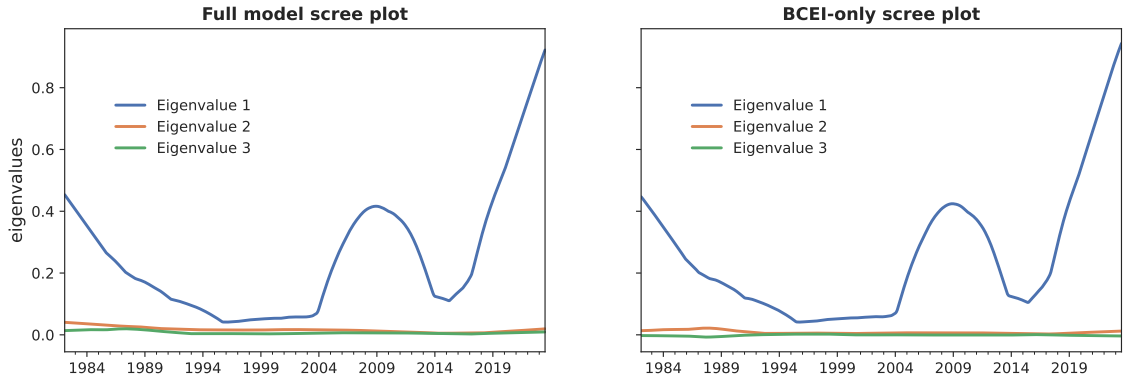

Figure 6: The three largest eigenvalues of the local sample covariance matrix from merged BCEI and CF data (left panel) and BCEI data only (right panel).

4.2 Results

We now present the key findings from our empirical analysis. The bandwidth for the tvHPCA was determined through cross-validation, resulting in a value of $b = 0.121$ (additional details can be found in the appendix). It is important to note that confidence intervals for impulse responses should be interpreted with caution due to the serial correlation of the extracted shocks in our application, which our simulations indicate may lead to some undercoverage.

[18] Borağan Aruoba (2020) similarly combines different data sources to construct a term structure of expectations (not revisions), but for the different purpose of linking asset prices to inflation expectations. We note that we obtain a smaller set of points in the term structure than Borağan Aruoba (2020), because for our purpose it is paramount that the expectations from different data sources are based on aligned information sets.

4.2.1 One perceived shock, highly correlated with inflation surprises

To determine the number of shocks, we apply the procedure described in Section 2.4, based on a local (in time) version of the test proposed by Onatski (2010).

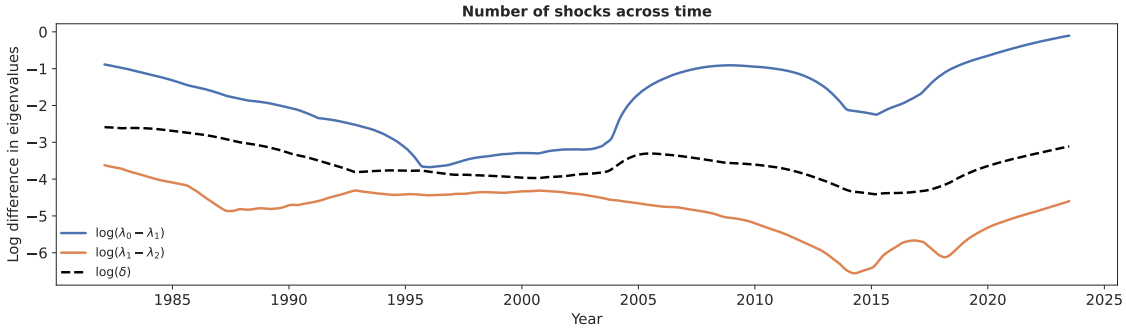

Figure 7: Results of the test for the number of factors developed in Onatski (2010) applied across time. The solid blue line depicts the log difference between the 1st largest eigenvalue and the "0" factors case across time; the orange solid line depicts the log difference between the 1st and 2nd largest eigenvalues. The dashed black line represents the critical values at each point in time.

From Figure 7 we can see clear evidence of 1 shock consistently across time. This is one of our main stylized facts. We then extract the shock by applying tvHPCA using the normalization in Assumption 2(a). To gain insight into the nature of the extracted shock, we try to relate it to actual data for the period under consideration. First, we find that the shock is highly correlated with the 3-month 3-month annualized rate of inflation (correlation 0.73) over the time period February 1982 to July 2023.[19] This could be interpreted as showing a fairly high correlation between the shock and the surprises from a model that forecasts inflation as being constant (e.g., at a given target). We then estimate and compute surprises from Stock and Watson (2007)'s Unobserved Component Stochastic Volatility (UCSV) model over the same

[19] 3-month 3-month inflation is calculated using seasonally adjusted headline CPI from U.S. Bureau of Labor Statistics (2023) as $[(I_t + I_{t-1} + I_{t-2})/(I_{t-3} + I_{t-4} + I_{t-5})]^4 - 1$, where I_t is the CPI index in month t. While monthly inflation rates would be theoretically more reflective of new information, they are more volatile and showed weaker correlation with the extracted inflation shock. This may be reflective of information frictions, whereby the inflation shock incorporates very recent but not exclusively the latest monthly inflation release.

period[20] for quarter-on-quarter annualized CPI inflation.[21] We find an even higher correlation between our shock and the surprises from this model (the correlation equals 0.81 over the whole sample, and 0.85 when only considering data up to 2020).

The high correlation between the shock and the surprises from the UCSV model can be seen in Figure 8. The figure also shows a change in pattern post-pandemic, with shocks of persistently larger magnitude than the surprises from the UCSV model. This could be interpreted as suggesting that agents in our expectation data kept underestimating the persistence of inflation in the post-pandemic period, and thus perceived a sequence of positive shocks, whereas the UCSV model more accurately characterized the persistence of inflation in the data.

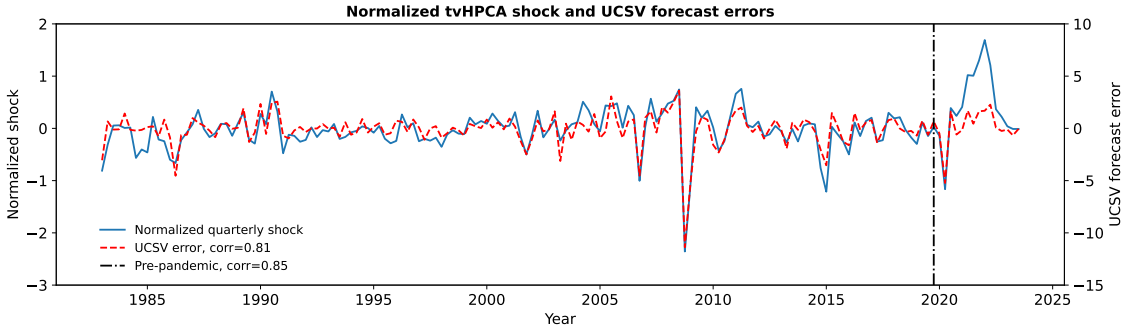

Figure 8: Perceived shock vs. surprises from Stock and Watson's (2007) model

4.2.2 Secular decrease in the perceived persistence of the effect of the shock

Below we show our estimates of the loadings across horizons and over time, obtained under the normalization 2(a) of a unit local standard deviation of the shock.

[20]The solution to the UCSV model is computed by Markov Chain Monte Carlo (MCMC) using a diffuse prior for the initial condition and $\gamma = 0.2$ as the sole model parameter. The Matlab code from Chan (2018) can be accessed at `https://joshuachan.org/code/code_spectest.html`. The surprises are the estimated model's residuals.

[21]The quarter-on-quarter annualized CPI inflation is also calculated using seasonally adjusted headline CPI from U.S. Bureau of Labor Statistics (2023) as $(I_q/I_{q-1})^4 - 1$, where I_q is the average of monthly CPI in a given quarter q.

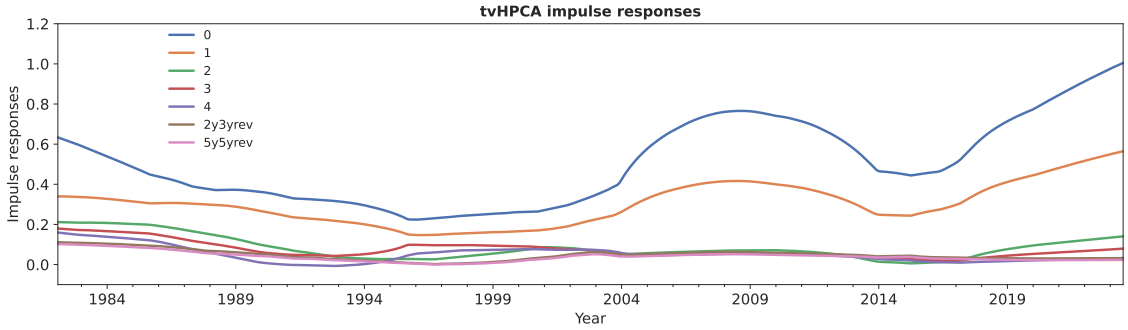

Figure 9: The graph presents the estimated impulse responses across the term structure of horizons and over time.

Figure 9 shows clear evidence of time variation in impulse responses across time. The impulse response function (capturing the dynamic effects of the shocks across different horizons) at a given point in time can be obtained as a vertical slice from the figure. These slices suggest time-varying shapes of the impulse response function, with the effects of the shock generally decreasing with the horizon. The apparent non-monotonicity at longer horizons in the 1990s is not statistically significant (see Figure 10, which plots 95% confidence intervals for selected horizons between 1993-1999).[22]

[22]Monotonicity requires that the 4th horizon impulse response lie between the 2nd and 5-year 5-year horizon impulse response. The point estimate for the 4th horizon impulse response is below that of the 5-year 5-year horizon in the early 1990s and exceeds that of the 2nd horizon in the late 1990s. However, the overlap in confidence intervals shows that these deviations are not statistically significant.

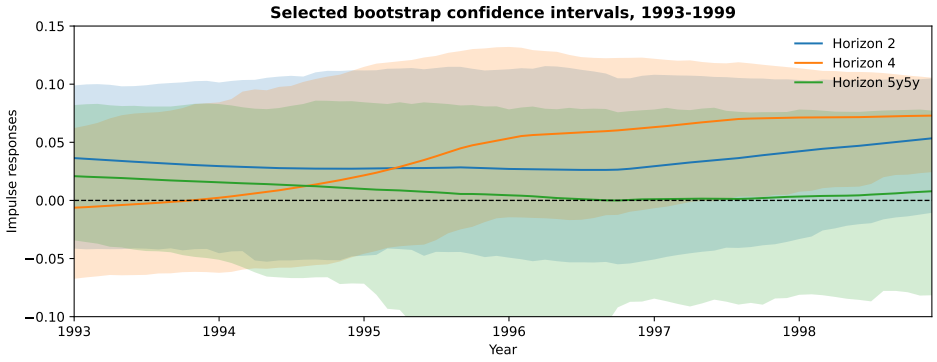

Figure 10: 95% bootstrap confidence intervals for impulse responses at selected horizons, 1993-1999.

Because of the unit-standard-deviation normalization for the shock used in the figure, the magnitude of the impulse responses reflects the volatility of expectation revisions at that horizon. We see that this volatility increased around the financial crisis of 2009 and at the end of our sample, which includes the Covid pandemic and the post-pandemic recovery. The difference between the impulse responses at a point in time contains information about the perceived persistence of the effect of the shocks, indicating for example quickly decaying responses to shocks around 2009.

Because the standard deviation of the shocks changes over time, the unit-standard deviation normalization has too many moving parts to enable a clear comparison of impulse responses over time. We thus recompute the impulse responses using the unit-impact normalization in Assumption 2(c) and plot in Figure 11 the impulse response functions for three selected times: during the Volker disinflation period of 1986, during the financial crisis of 2009 and during the high-inflation period of 2022. The figure also reports bootstrap confidence intervals at each point in time, computed as described in Section 2.6.

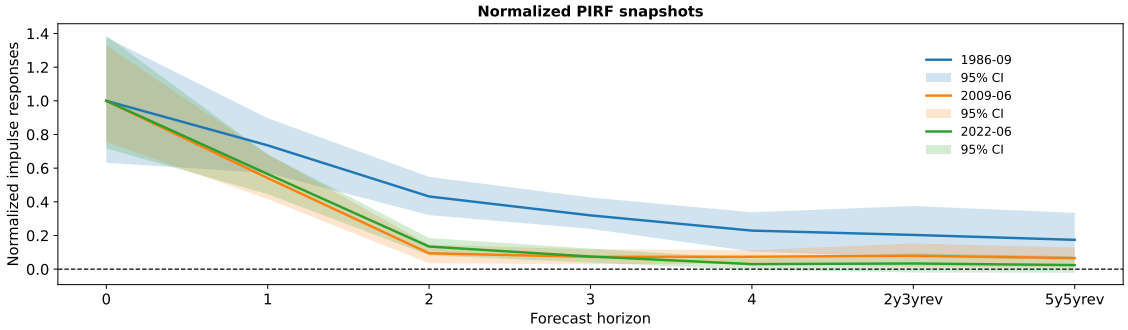

Figure 11: Perceived impulse response functions during the Volker disinflation, the financial crisis of 2009 and the 2022 high-inflation period

We see that during the Volker disinflation years, long-term inflation expectations were deanchored, in the sense that the perceived persistence of the effect of the shock remained high at long horizons. While in 2009 and 2022 the short-term inflation volatility was higher (as reflected by the large impulse responses for the 0th horizon in Figure 9), long-term inflation expectations remained anchored. In fact, the long-term impulse response has decreased over time for these snapshots.[23] It is worth noting that, although long-term inflation expectations were at risk of becoming deanchored in 2022, our data imply a historically low persistence of the shock at long horizons (subject to the discussed caveats of nonparametric estimation at the sample boundaries).

The finding of time-varying (and rapidly decaying) impulse responses also suggests that agents do not use Stock and Watson (2007)'s UCSV model to produce forecasts, in spite of our finding in the previous section that our extracted shock is highly correlated with the surprises from this model. This is because the UCSV model implies a flat impulse response function (as can be deduced from equation (2) and the discussion thereafter, the UCSV's model-implied impulse response equals 1 at all horizons), which we do not observe in the data.

[23]This is consistent with the finding of Stock and Watson (2007) - and extends it to the current period - that transitory shocks to inflation have become more volatile/relevant while permanent shocks to inflation have become less volatile/relevant.

4.2.3 A possible narrative about the 2022 high-inflation episode

Putting together the findings in the previous two sections, we can provide a possible narrative for what happened during the post-pandemic high-inflation episode (again subject to the caveat of possible increased bias at the sample boundaries).

In general, our method can provide an answer to the policy-relevant question: if we see a large change in long-term inflation expectations, is it because agents perceived a large shock or because they expected the shock to persist (i.e., a notion of deanchoring)? The method can answer this question by disentangling the two latent sources. The two subpanels of Figure 12 plot the perceived impulse response functions at two points in time during 1986 and 2022 (left panel) and the full time series of the extracted shock (right panel, smoothed using a 12-month moving average).

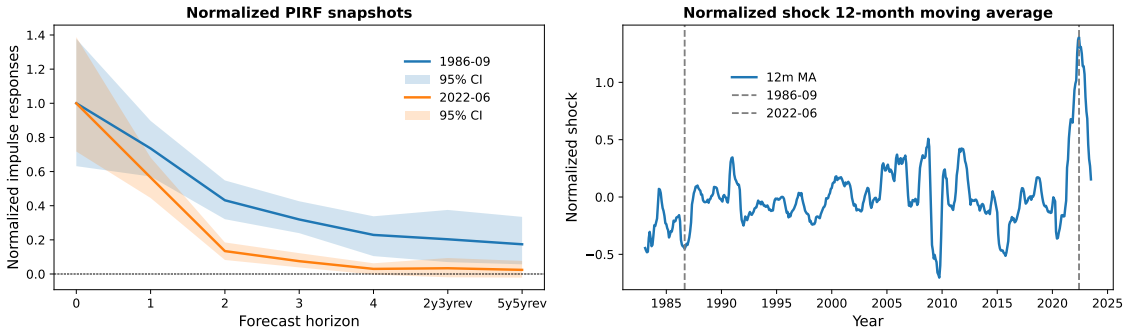

Figure 12: Perceived impulse response functions and corresponding shocks: 1986 vs. 2022

The figure reveals a clear contrast between the two dates: during 1986 agents perceived a shock of normal magnitude by historical standards, but believed the shock to be highly persistent (deanchoring), whereas in 2022 the perceived shock was unprecedentedly large, but agents believed that its effects would essentially disappear within a year (anchoring).

5 Conclusion

This paper demonstrates how leveraging the typically short horizon dimension present in various expectations datasets, combined with a focus on expectation revisions, facilitates the construction of novel empirical measures of how beliefs about shocks and impulse responses evolve over time. Central to this approach is modeling these revisions using a time-varying factor model applied to panel data across different horizons and periods. This framework yields estimates of shocks (the factors) and potentially time-varying impulse responses (the loadings). Our nonparametric method, based on minimal assumptions, exhibits strong performance in extracting shocks and impulse responses even in small samples, and remains robust to heteroskedasticity and serial correlation. However, it is important to recognize that this robustness in estimation currently comes at the expense of a fully developed theoretical foundation for inference and selection of the number of factors in a finite-sample setting.

Our method's versatility allows it to address various economic questions and adapt to different types of expectations data. While our empirical application focused on aggregate expectations in a balanced panel, the small-sample nature of our approach opens possibilities for analyzing individual-level expectations, potentially revealing heterogeneity in beliefs about shocks and their dynamic effects.

Moreover, we note the potential to explore an additional dimension of the data that we did not exploit in this paper: expectations related to multiple variables. This extension could enhance our understanding of the expectation formation process by revealing whether agents respond to the same shocks when forecasting different variables and whether their beliefs align with economic theory. For example, analyzing forecasts of both inflation and output could reveal insights into forecasters' beliefs about supply and demand shocks. Similarly, forecasts of interest rates and inflation could shed light on perceived policy rules and the effectiveness of forward guidance. These applications present promising avenues for future research.

Appendix A: Assumptions

The model for the expectation revisions in (1) can be written in vector notation as:

$$\underset{H\times 1}{X_t} = \underset{H\times r}{\Lambda_t}\ \underset{r\times 1}{F_t} + \underset{H\times 1}{e_t}\,, \tag{15}$$

where $\Lambda_t = (\lambda_{1,t}, \ldots, \lambda_{H,t})'$ is the matrix of loadings (i.e., the perceived impulse responses), F_t is the vector of common factors (i.e., the perceived shocks) and e_t is the vector of idiosyncratic errors. Our method is based on *local* (in time) PCA, and thus the objects of interest are *local* covariance matrices, defined as follows:

$$\Sigma_t = E\left[X_t X_t'\right], \qquad \Sigma_{F,t} = E\left[F_t F_t'\right], \qquad \Sigma_{e,t} = E\left[e_t e_t'\right]. \tag{16}$$

Our method relies on the following assumptions, which we group into different types to facilitate the discussion of their practical implications.

A.1 Shocks and idiosyncratic errors

Assumption 1 (Shocks and idiosyncratic errors):

(a) The shocks F_t have unconditional mean zero.

(b) The shocks F_t are cross-sectionally independent but can be serially correlated (stationary). Moreover, $cov(e_{h,t}, F_{jt-k}) = 0$ for any h, j, t and k.

(c) The errors e_t can be serially correlated (stationary), and $e_t \sim (0, \Sigma_{e,t})$, where $\Sigma_{e,t}$ is a diagonal matrix with uniformly bounded eigenvalues for all $t = 1, \ldots, T$.

Comments. Assumption 1(a) requires shocks to have unconditional mean zero, which is plausible since we analyze expectation revisions; a non-zero mean would suggest forecast bias, unlikely in typical expectations data.[24] Assumption 1(b) is

[24] A constant mean can be accommodated by demeaning $X_{h,t}$ and restoring the mean after extracting shocks.

the standard PCA assumption that factors are independent of each other and errors. Note that we do not require serial independence. Our simulations show that serial correlation in shocks or errors does not affect the performance of tvHPCA in bias or correlation measures. The main requirement, Assumption 1(c), states errors are uncorrelated across horizons. Although examples often assume no idiosyncratic errors, these can arise from survey respondent variation, rounding, or dataset merging, which are likely uncorrelated across horizons. Simulations indicate that small cross-sectional correlation has little impact, and HPCA outperforms PCA even with some correlation. When the number of horizons H is large, PCA's robustness to error correlation is well-established (see Bai and Wang (2016)).

Practical Implications. In practical terms, Assumption 1 implies that, although serial correlation in shocks and errors does not impede the methodology's ability to recover shocks and impulse responses, error correlation across horizons introduces some bias in the estimates of impulse responses in very small samples. The simulations in this paper suggest that the extraction of shocks is less affected by violation of the assumption of cross-sectionally uncorrelated errors and that, in any case, HPCA continues to outperform PCA even in the presence of cross-sectional correlation in errors.

A.2 Normalizations

As in standard factor models, there is a rotational indeterminacy in the identification of shocks and impulse responses.[25] This can be resolved by imposing a normalization either on the shocks or the impulse responses, depending on the objectives of the analysis. We consider the following examples of normalizations.

Assumption 2 (Normalizations):

(a) *Unit-local shock variance normalization:* Set $\Sigma_{F,t} = I_r$, where I_r is the identity matrix.

[25]For any $r \times r$ nonsingular matrix A_t it follows that $\lambda'_{h,t}F_t = (A_t^{-1}\lambda_{h,t})'(A'_tF_t)$ and therefore shocks and impulse responses are not separately identified.

(b) *Unit-effect normalization:* Set $\Lambda_t'\Lambda_t = I_r$, such that the shocks have unit effect on all horizons.

(c) *Unit-impact normalization:* Set $\Lambda_t[1,h] = 1$ for $h = 1, \cdots, H$, such that the shocks have unit effect only on the first horizon (i.e., on impact).

Practical Implications. The practical implication of Assumption 2 is that one must choose whether to normalize the shocks or the impulse responses based on the primary focus of the analysis. If the emphasis is on the impulse responses, the shocks can be normalized to have a unit local variance (Assumption 2(a)), allowing the impulse responses to reflect the dynamic effects of a one-standard-deviation shock. Conversely, if the goal is to recover the shocks, the impulse responses can be normalized so that the shock magnitudes are interpretable (Assumption 2(b)). Alternatively, one can impose a unit-impact normalization on the impulse responses (Assumption 2(c)), which rescales the shocks to produce a unit effect on the first horizon at each point in time. This normalization facilitates comparisons of impulse response functions over time.

A.3 Time-varying impulse responses and nonparametric estimation

The time-varying impulse responses are assumed to be deterministic functions of time:

$$\lambda_t = \lambda(t/T) \qquad \text{and} \qquad \Lambda_t = \Lambda(t/T). \tag{17}$$

Such rescaling is common in nonparametric estimation (see e.g. Robinson, 1989, Cai, 2007). The idea is that, as the number of observations increases in the rescaled time framework, we "observe" the process on an increasingly dense grid on the unit interval, and letting $T \to \infty$ allows for the impulse responses to be consistently estimated under infill asymptotics, see e.g. Motta et al. (2011), Su and Wang (2017). Note that, since the $\lambda_{h,t}$ in (1) is time-varying, $X_{h,t}$ is no longer stationary, but locally stationary in the sense of Dahlhaus et al. (2019), i.e., behaving in an approximately

stationary manner within short time periods. With the new notation, the locally stationary model reads:

$$X_{t,T} = \Lambda(t/T)\, F_t + e_t, \tag{18}$$

where time variation in impulse responses gives rise to a triangular array $X_{t,T}$. This framework allows us to analyze the dynamics of $X_{t,T}$ locally by using a stationary approximation[26] given by:

$$X_t(u) = \Lambda(u)\, F_t + e_t, \tag{19}$$

where $X_t(u)$ is a locally stationary equivalent of $X_{t,T}$ and where $\Lambda(t/T) \approx \Lambda(u)$. Note that $X_t(u)$ is not observed in practice but is used as a theoretical construct. For ease of exposition in the paper we simply write X_t.

Our method is based on the nonparametric estimator of the local covariance matrix of X_t reported in equation (4). We make the following assumptions.

Assumption 3 (Time-varying impulse responses and kernel estimation):

(a) $\lambda_{h,t} = \lambda_h(t/T)$, $t = 1, \ldots, T$, where for $h = 1, \ldots, H$, $\lambda_h(\cdot) : [0,1] \to \mathbb{R}$ is an unknown piece-wise continuous function of the rescaled time t/T. $\lambda_h(\cdot) : [0,1] \to \mathbb{R}$ is twice continuously differentiable for any $h = 1, \ldots, H$.

(b) For all $u \in (0,1)$ it holds that $rank(\Lambda(u)) = r$, where $\Lambda(u)$ is defined in eq.(19).

(c) The kernel function $K : \mathbb{R} \to \mathbb{R}^+$ is a symmetric continuously differentiable probability density function with compact support $[-1,1]$ normalized such that $\int K(z)dz = 1$.

(d) As $T \to \infty$ $b \to 0$, such that $Tb \to \infty$.

Comments. Assumption 3 is similar to the assumptions in Su and Wang (2017). Assumption 3(a) requires the time-varying impulse responses to be smooth (i.e., continuously differentiable) functions of time. Assumption 3(b) states that the number

[26]The idea is to approximate $X_{t,T}$ by $X_{t,T} = X_t(t/T) \approx X_t(u)$, where $X_t(u)$ is a stationary equivalent of $X_{t,T}$.

of shocks is fixed across time. The assumption could be relaxed, but this would come at the cost of losing interpretability and the ability to label the shocks across time. Assumption 3(c) is the standard assumption in the nonparametric literature, applied here to estimation of the local covariance matrix, requiring the chosen kernel to be a symmetric probability density function. In the paper we use the Epanechnikov kernel $K(x) = 0.75(1 - x^2)\mathbb{1}\{|x| \leq 1\}$, rescaled when necessary to ensure consistency even on the boundary points, where only the data on one side are available.[27] Assumption 3(d) states typical conditions on the bandwidth b that ensure that the local covariance matrix can be consistently estimated locally in time, see e.g. Motta et al. (2011) and Su and Wang (2017).

Practical implications. Assumption 3 has several practical implications. First, to assess whether the number of shocks remains constant over time, we use a local version of Onatski (2010) (detailed in section 2.4); while not a formal finite-sample test, it can suggest caution if time variation is indicated. Second, since shocks and impulse responses are identified only up to a sign,[28] maintaining continuity in impulse responses requires additional steps. The tvHPCA algorithm achieves this through two sub-algorithms: Algorithm A extracts shocks and time-varying responses, and Algorithm B ensures their continuity. Third, a sufficiently large time dimension T is needed to recover time-varying impulse responses; with smaller T, the method simplifies to the time-invariant case in Zhang et al. (2022). Finally, continuity of impulse responses and a stable number of shocks are crucial for interpretation: if shocks appear or vanish abruptly, it becomes difficult to identify whether they are the same or different shocks.

[27] This is equivalent to applying the boundary kernel for the boundary regions, see e.g. Li and Racine (2023) and is the same as the boundary kernel applied in Su and Wang (2017). The rescaling is necessary to achieve consistency of the corresponding estimates on the boundary points by ensuring that the first moment of the kernel is always normalized to 1.

[28] Normalization for scale is addressed by Assumption 2.

A.4 Local incoherence

The final assumption is known as the incoherence condition, a standard concept in the matrix completion literature (see e.g., Candès and Recht, 2009, Zhang et al., 2022). This condition ensures that the information contained in the row and column spaces of the covariance matrix is not concentrated in too few rows or columns. This is important because it enables the original high-dimensional matrix to be approximated by a lower-dimensional one.

Given the time-varying nature of our approach, we need to apply this incoherence condition locally, i.e., a "local incoherence" assumption.

Assumption 4 (Local Incoherence): There exists a constant c such that the following holds for $t = 1, \ldots, T$:

$$\max_{1 \leq h \leq H} \|\nu_h \Lambda_t (\Lambda_t' \Lambda_t)^{-1/2}\|_2^2 \leq c, \tag{20}$$

where ν_h is the h-th standard basis of appropriate dimension with h-th coordinate equal to 1 and other coordinates 0s and, for a vector v, $\|v\|_2$ denotes its l_2-norm.

Comments. Condition in (20) is a slightly modified version of the incoherence condition found in e.e.g Candès and Recht (2009), Zhang et al. (2022) to ensure it holds under all normalizations stated in our Assumption 2. Specifically, we re-normalize the loadings such that they are orthonormal regardless of the chosen normalization. While Assumption 4 is high-level, it relates to the more familiar concept of "strong factors" (see, e.g., the discussion in Agarwal et al., 2023) and to the "pervasiveness" assumption found in the economics literature, see e.g. Onatski (2012), and Fan et al. (2013). These assumptions ensure that the common component can be discerned from the idiosyncratic component of the sample variance. For example, under the unit-local shock variance normalization in Assumption 2(a), the condition corresponds to the pervasiveness assumption $\psi_{\min}(\Lambda_t' \Lambda_t / H) > c > 0$, where $\psi_{\min}(M)$ denotes the smallest eigenvalue of matrix M.

Practical implications. Assumption 4 suggests caution with "weak factors", where many or all responses are near zero - possibly due to infrequent revisions or too

small a bandwidth for covariance estimation. In Section 3, we examined a worst-case scenario with low signal-to-noise ratio, many zeros, and rapidly decaying responses, offering practical guidance on diagnosing and understanding such violations.

Appendix B: Bandwidth selection

This appendix presents the details of the cross-validation used to determine the optimal bandwidth b for nonparametric estimation of the sample covariance matrix. Given that the estimation is done via local singular value decomposition followed by several steps of refinements, the bias-variance trade-off necessary to discuss a theoretically optimal bandwidth is non-standard. We therefore follow Su and Wang (2017) and use a data-driven way of selecting an empirically optimal bandwidth by a version of cross-validation (CV). Specifically, we select the empirically optimal bandwidth $\widehat{b}^*$ by solving the following minimization problem:

$$\min_b CV(b) = \frac{1}{Tp} \sum_{h=1}^{p} \sum_{s=1}^{T} \left[X_{h,s} - \widehat{\lambda}_{h,s}^{(-s)} \widehat{F}_t^{(-s)} \right]^2, \tag{21}$$

where $\widehat{\lambda}_{h,s}^{(-s)}$ and $\widehat{F}_t^{(-s)}$ are the versions of $\widehat{\lambda}_{h,s}$ and $\widehat{F}_t$ respectively with the s-th time series removed. In the context of our application, the graph of the CV distance, defined in (21), is presented below with the resulting optimal bandwidth $\widehat{b}^* = 0.121$.

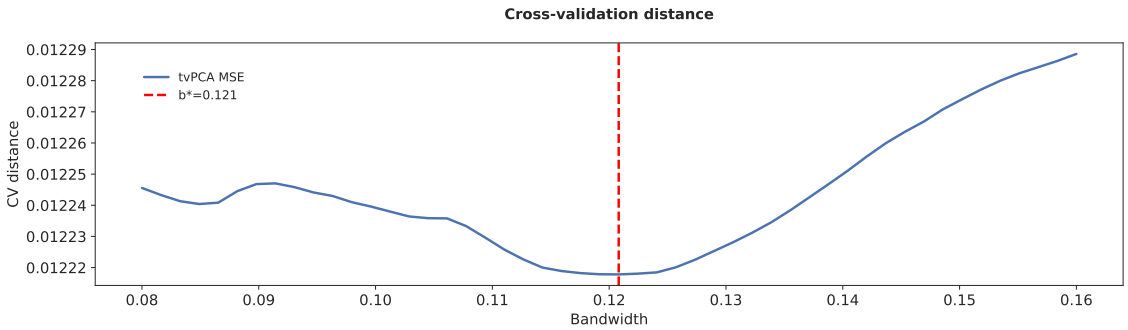

Figure 13: Plot of CV distance (eq. 21) in the empirical application, with red dashed line indicating optimal bandwidth.

During the preparation of this work the authors used ChatGPT in order to help edit individual paragraphs. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Agarwal, A., Agarwal, A., and Vijaykumar, S. (2023). Synthetic combinations: A causal inference framework for combinatorial interventions. *Advances in Neural Information Processing Systems*, 36:19195–19216.

Anderson, T. and Rubin, H. (1956). Statistical inference in. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability: Held at the Statistical Laboratory, University of California, December, 1954, July and August, 1955*, volume 1, page 111. Univ of California Press.

Aruoba, S. B. (2020). Term structures of inflation expectations and real interest rates. *Journal of Business & Economic Statistics*, 38(3):542–553.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Bai, J. and Ng, S. (2019). Rank regularized estimation of approximate factor models. *Journal of Econometrics*, 212(1):78–96.

Bai, J., Ng, S., et al. (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics*, 3(2):89–163.

Bai, J. and Wang, P. (2016). Econometric analysis of large factor models. *Annual Review of Economics*, 8(1):53–80.

Bianchi, F., Ludvigson, S. C., and Ma, S. (2022). Belief distortions and macroeconomic fluctuations. *American Economic Review*, 112(7):2269–2315.

Blanchard, O. J., L'Huillier, J.-P., and Lorenzoni, G. (2013). News, noise, and fluctuations: An empirical exploration. *American Economic Review*, 103(7):3045–3070.

Blue Chip Economic Indicators (2023). Aspen Publishers: Blue Chip Economic Indicators. https://www.wolterskluwer.com/en/solutions/blue-chip (accessed August 21, 2023).

Borağan Aruoba, S. (2020). Term structures of inflation expectations and real interest rates. *Journal of Business & Economic Statistics*, 38(3):542–553.

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics*, 136(1):163–188.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found Comput Math*, 9:717–772.

Carvalho, C., Eusepi, S., Moench, E., and Preston, B. (2023). Anchored inflation expectations. *American Economic Journal: Macroeconomics*, 15(1):1–47.

Chamberlain, G. (1983). Funds, factors, and diversification in arbitrage pricing models. *Econometrica: Journal of the Econometric Society*, pages 1305–1323.

Chan, J. C. (2018). Specification tests for time-varying parameter models with stochastic volatility. *Econometric Reviews*, 37(8):807–823.

Chen, Z. and Leng, C. (2015). Local linear estimation of covariance matrices via cholesky decomposition. *Statistica Sinica*, pages 1249–1263.

Choi, I. (2012). Efficient estimation of factor models. *Econometric theory*, 28(2):274–308.

Coibion, O. and Gorodnichenko, Y. (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8):2644–2678.

Crump, R. K., Eusepi, S., Moench, E., and Preston, B. (2023). The term structure of expectations. In Bachmann, R., Topa, G., and van der Klaauw, W., editors, *Handbook of Economic Expectations*, pages 507–540. Academic Press.

Dahlhaus, R., Richter, S., and Wu, W. B. (2019). Towards a general theory for nonlinear locally stationary processes. *Bernoulli*, 25(2).

Del Negro, M. and Otrok, C. (2008). Dynamic factor models with time-varying parameters: measuring changes in international business cycles. *FRB of New York Staff Report*, (326).

Diebold, F. X. and Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of econometrics*, 130(2):337–364.

Diebold, F. X., Piazzesi, M., and Rudebusch, G. D. (2005). Modeling bond yields in finance and macroeconomics. *American Economic Review*, 95(2):415–420.

Enders, Z., Kleemann, M., and Müller, G. J. (2021). Growth expectations, undue optimism, and short-run fluctuations. *Review of Economics and Statistics*, 103(5):905–921.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(4):603–680.

Federal Reserve Bank of Cleveland (2023). Cleveland Fed Inflation Expectations Series. https://www.clevelandfed.org/indicators-and-data/inflation-expectations (accessed August 21, 2023).

Florescu, L. and Perkins, W. (2016). Spectral thresholds in the bipartite stochastic block model. In *Conference on Learning Theory*, pages 943–959. PMLR.

Gertler, M. and Karadi, P. (2015). Monetary policy surprises, credit costs, and economic activity. *American Economic Journal: Macroeconomics*, 7(1):44–76.

Gürkaynak, R. S., Sack, B., and Swanson, E. (2005). The sensitivity of long-term interest rates to economic news: Evidence and implications for macroeconomic models. *American economic review*, 95(1):425–436.

Herbst, E. and Winkler, F. (2021). The factor structure of disagreement. Available at SSRN: https://ssrn.com/abstract=3872757.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

Kim, H. (2023). Quantifying the sources of forecaster disagreement. Available at SSRN: https://ssrn.com/abstract=4659361.

Kuttner, K. N. (2001). Monetary policy surprises and interest rates: Evidence from the fed funds futures market. *Journal of Monetary Economics*, 47(3):523–544.

Li, Q. and Racine, J. S. (2023). *Nonparametric econometrics: theory and practice*. Princeton University Press.

Maldonado, J. and Ruiz, E. (2021). Accurate confidence regions for principal components factors. *Oxford Bulletin of Economics and Statistics*, 83(6):1432–1453.

Motta, G., Hafner, C. M., and von Sachs, R. (2011). Locally stationary factor models: Identification and nonparametric estimation. *Econometric Theory*, 27(6):1279–1319.

Nakamura, E. and Steinsson, J. (2018). High-frequency identification of monetary non-neutrality: The information effect. *The Quarterly Journal of Economics*, 133(3):1283–1330.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642.

Plagborg-Møller, M. (2019). Bayesian inference on structural impulse response functions. *Quantitative Economics*, 10(1):145–184.

Robinson, P. M. (1989). *Nonparametric estimation of time-varying parameters*. Springer.

Stock, J. H. and Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Stock, J. H. and Watson, M. W. (2006). Forecasting with many predictors. *Handbook of economic forecasting*, 1:515–554.

Stock, J. H. and Watson, M. W. (2007). Why has us inflation become harder to forecast? *Journal of Money, Credit and banking*, 39:3–33.

Stock, J. H. and Watson, M. W. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of macroeconomics*, volume 2, pages 415–525. Elsevier.

Su, L. and Wang, X. (2017). On time-varying factor models: Estimation and testing. *Journal of Econometrics*, 198(1):84–101.

Swanson, E. T. (2021). Measuring the effects of federal reserve forward guidance and asset purchases on financial markets. *Journal of Monetary Economics*, 118:32–53.

U.S. Bureau of Labor Statistics (2023). Consumer Price Index for All Urban Consumers: All Items in U.S. City Average. https://www.bls.gov/cpi/ (accessed August 21, 2023).

Zhang, A. R., Cai, T. T., and Wu, Y. (2022). Heteroskedastic pca: Algorithm, optimality, and applications. *The Annals of Statistics*, 50(1):53–80.